

**Impact of bias in evolutionary mechanisms: migration among
subdivided population and gene conversion among multigene family**

Shuhei Mano

Graduate School of Natural Sciences, Nagoya City University, Nagoya 467-8501, Japan

Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi 332-0012, Japan

Address for correspondence: Graduate School of Natural Sciences, Nagoya City University,
1 Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya 467-8501, Japan;
E-mail: mano@nsc.nagoya-cu.ac.jp; Tel/Fax: +81-(0)52-872-5821

Abstract

Two models of bias in evolutionary mechanism is presented: one of bias in migration in Wright's island model and the other of bias in gene conversion among members of a multigene family in a panmictic population. The models have an identical diffusive limit for a large population size, where allele frequencies in each subpopulation of the n -island model are identical to allele frequencies in each locus of the n -gene family. The probability of fixation of a new mutant throughout the total population/gene family is obtained by the diffusion process. It is shown that the deviation of the probability of fixation from that without bias is proportional to one plus the number of migrants/converted alleles. For the island model with allele-dependent migration, an analogue of the coalescent genealogy, which we call the ancestral bias graph, is introduced. We present a recursion formula that can be used to compute the probability of obtaining a given sample. We apply our formula to data set of mouse histone gene family. It is suggested that bias in ectopic gene conversion, whose magnitude is orders of magnitude larger than that in allelic gene conversion, can be maintained in a population. Recently, evidence of the large impact of biased gene conversion on gene substitution is accumulating. The fact that the diffusion models for conversion bias and for migration bias are identical suggests that migration bias can also have large impact on genomic polymorphism and divergence in subdivided populations.

Keywords: biased gene conversion, biased migration, diffusion model, ancestral process, biased voter model

1. INTRODUCTION

Evolutionary mechanisms are frequently biased. Because a slight bias in an evolutionary mechanism could cause result in an apparent natural selection in neutral loci, a quantitative understanding effect of biases in evolutionary mechanisms is necessary for understanding molecular evolutionary mechanisms.

Recently, evidence of bias in gene conversion, which is a recombination associated molecular drive that favors AT to GC mutations, is accumulating [1, 2]. Nagylaki [3] showed that the bias in allelic conversion is equivalent to directional selection. Regions of a genome that evolve rapidly are generally regarded as being under strong positive selection. However, it was shown that many protein coding changes in the fastest changing genes of human genome are not a result of selection operating on the genes; rather, they result from biased fixation of AT to GC mutations [4]. It was also suggested that the ectopic gene conversion among members of multigene families under concerted evolution [5] is also biased. In histone paralogous genes obtained from humans and mice, it was found that gene copies that belong to subfamilies with very similar sequences (presumably undergoing gene conversion) have a high GC content than unique gene copies (presumably not undergoing gene conversion) [6].

In human populations, migrants are rarely a random sample of their source population. It was demonstrated that patrilocal populations, where men tend to stay at their birthplace while women move to their husband's birthplace, exhibit greater Y-chromosomal than mitochondria DNA genetic differentiation, and in matrilocal populations, where women tend to stay at their birthplace, the converse trend is observed [7]. In addition, differences between the social organization difference of herders and agriculturists could enhance the genetic differentiation [8]. It is fairly reasonable to assume that a migration rate depends on allelic types. If so, neutral evolution of subdivided population might not be seen as being under neutral evolution. The difference of the migration rates here is not necessarily intended to be a result of the difference in biological functions associate with the allelic

types. Consider a population founded as a mixture of two ancestral populations with very differentiated genetic backgrounds. If one ancestral population have higher tendency of migration than the other ancestral population, alleles specific to the former population should have higher migration rate than alleles specific to the latter population.

Hence, bias in evolution is an intriguing issue. Nevertheless, a few efforts have been devoted to modeling such biases. In this study, we will discuss biases in two evolutionary processes: gene conversion among members of multigene family in a panmictic population and Wright's island model [9] with allele-dependent migration. We will present continuous-time Moran models for these biases. The diffusive limits of these models are identical, where allele frequencies in each subpopulation of the n -island model are identical to allele frequencies in each locus of the n -gene family. The probability of fixation of a new mutant throughout the total population/gene family is obtained from the diffusion process. Then, we describe the island model with biased migration in terms of a biased voter model [10]. By using a duality, we obtain a random graph, which is analogous to the coalescent genealogy [11, 12, 13]. We call this graph the ancestral bias graph. We present a recursion that can be used to approximate the probability of obtaining a given sample by simulating backwards along sample path of the ancestral bias graph. In the strong migration/conversion limit [14], the diffusion model is equivalent to that for directional selection, and thus, the ancestral bias graph reduces to the ancestral selection graph [15]. By using the formula, we demonstrate quantifying biases in gene conversion among multigene families.

2. FORMULATION

Consider a subdivided population consists of n demes, where each deme is occupied by N haploid individuals and all pairs of demes can exchange migrants symmetrically (n -island model). The population evolves according to a continuous-time Moran model with revertible mutation. The sizes of the demes are kept constant by migrations. We

will limit the discussion to two allelic types, A_1 and A_2 , and assume that the mutation between the two alleles are symmetric. The Moran model is a continuous-time Markov process in which individuals produce one offspring at a time. The type of offspring will be chosen according to the mutation process. The offspring will then replace an individual who is chosen at random from the same deme or the other demes. The offspring may replace its own parent. The replaced individual is removed from the population, and thus, the deme size is kept constant. We assume that an individual reproduces at a rate of λ_0 and replaces an individual from the same deme. In addition, in the migration process, alleles A_1 and A_2 replace an individual in the other demes at rates $\lambda_0\xi_1$ and $\lambda_0\xi_2$, respectively. The offspring will have the same type as the parent with probability $1 - u$ and will have the other type with probability u , $u \in [0, 1]$. We set $\xi_1 = m(1 - b)/(n - 1)$ and $\xi_2 = m(1 + b)/(n - 1)$ with $0 \leq b < 1$ and $m \geq 0$. The migration event is biased if $b > 0$. The state of the population at time t can be represented as a continuous-time Markov chain $\mathbf{Z}(t) = (Z_i(t))$, $i = 1, 2, \dots, n$, where $Z_i(t)$ is the number of individuals of type A_1 in the i -th deme at time t . If $\mathbf{Z}(t) = \mathbf{z}$, $z_i = 0, 1, \dots, N$, the transition to $\mathbf{z} + \mathbf{e}_i$, where \mathbf{e}_i , $i = 1, 2, \dots, n$ is the unit vector, occurs at a rate of

$$(2.1) \quad \begin{aligned} & \lambda_0(z_i + \xi_1 \sum_{k(\neq i)} z_k) \frac{N - z_i}{N} (1 - u) \\ & + \lambda_0\{(N - z_i) + \xi_2 \sum_{k(\neq i)} (N - z_k)\} \frac{N - z_i}{N} u. \end{aligned}$$

and the transition to $\mathbf{z} - \mathbf{e}_i$ occurs at a rate of

$$(2.2) \quad \begin{aligned} & \lambda_0\{(N - z_i) + \xi_2 \sum_{k(\neq i)} (N - z_k)\} \frac{z_i}{N} (1 - u) \\ & + \lambda_0(z_i + \xi_1 \sum_{k(\neq i)} z_k) \frac{z_i}{N} u, \end{aligned}$$

Thus, $\mathbf{Z}(t)$ is a n -dimensional birth and death process with nonlinear birth and death rates. We consider the limiting diffusion approximation of the model described above.

We set $\lambda_0 = N/2$ and assume that $Nu \rightarrow \theta$ and $Nm \rightarrow \gamma$ as $N \rightarrow \infty$. $X_i(t)$ denotes

the fraction of genes of type A_1 in the i -th deme in the limiting process at time t . The generator of the diffusion process $\{\mathbf{X}(t); t \geq 0\}$ in an n -dimensional cube $[0, 1]^n$ is

$$(2.3) \quad L = L_0 - bL_1,$$

where

$$\begin{aligned} L_0 &= \sum_{i=1}^n \frac{x_i(1-x_i)}{2} \frac{\partial^2}{\partial x_i^2} + \frac{n\gamma}{2(n-1)} \sum_{i=1}^n (\bar{x} - x_i) \frac{\partial}{\partial x_i} \\ &\quad + \frac{\theta}{2} \sum_{i=1}^n (1-2x_i) \frac{\partial}{\partial x_i}, \\ L_1 &= \frac{\gamma}{2(n-1)} \sum_{i=1}^n \left\{ (1-2x_i) \sum_{j(\neq i)} x_j + (n-1)x_i \right\} \frac{\partial}{\partial x_i}. \end{aligned}$$

Here, \bar{x} is the arithmetic mean of \mathbf{x} . The limiting diffusion also appears as the limit of the discrete-time Wright-Fisher model, but the continuous-time Moran model is suitable for deriving the genealogical process.

Coincidentally, the diffusive limit for the n -island model is equivalent to that for a model of gene conversion among members of a multigene family consisting of n unlinked genes, where $X_i(t)$ denotes the fraction of genes of type A_1 in the i -th locus of the n -gene family. We now consider a monoecious panmictic population consists of N diploid individuals. The population evolves according to a continuous-time Moran model with revertible mutation. The Moran model is a continuous-time Markov process in which haplotypes, whose alleles at each locus are chosen randomly from the population, produce one offspring at a time. The type of the offspring will be chosen according to the mutation and the conversion process. The offspring will then replace a haplotype chosen at random from the population. The offspring may replace its own parent. The replaced individual is removed from the population, thus, the population size is kept constant. We assume that a haplotype reproduces at a rate λ_0 and replace a haplotype of the population. Let c be the rate at which a gene is converted by any one of the other $n-1$ genes with equal likelihood. Only a subset of the total conversion events involve different alleles. Among such conversion

events involving different alleles, let $(1+b)/2$ be the fraction of these event that result in allele A_1 being converted by allele A_2 , and similarly, let $(1-b)/2$ be the fraction of the events that result in allele A_2 being converted by allele A_1 , $0 \leq b < 1$ [3, 17]. We call the conversion event is biased if $b > 0$. The rate that an allele A_1 (A_2) is converted by an allele A_2 (A_1) is $c(1+b)/(n-1)$ ($c(1-b)/(n-1)$). For example, when $n = 3$, a haplotype $A_1A_1A_2$ changes to $A_1A_2A_2$, $A_2A_1A_2$, $A_1A_1A_1$ at rates $c(1+b)/2$, $c(1+b)/2$, and $c(1-b)$, respectively. For each locus, the offspring will have the same allelic type as the parent with probability $1-u$ and will have the other type with probability u . We present haplotypes using binaries, where the i -th digit is 1 and 0 when the locus is occupied by alleles A_1 and A_2 , respectively. The state of the population at time t can be represented as a continuous-time Markov chain $\mathbf{W}(t) = (W_\alpha(t))$, where $W_\alpha(t)$ is the number of haplotypes of type α in the population at time t . If $\mathbf{W}(t) = \mathbf{w}$, $w_\alpha = 0, 1, \dots, 2N$, the transition to $\mathbf{w} + \mathbf{e}_\alpha$ occurs at a rate of

$$(2.4) \quad \lambda_0 w_\alpha \frac{2N - w_\alpha}{2N} (1 - \sum_{\beta(\neq \alpha)} Q_{\alpha\beta}) + \lambda_0 \sum_{\beta(\neq \alpha)} w_\beta \frac{2N - w_\alpha}{2N} Q_{\beta\alpha}$$

and the transition to $\mathbf{w} - \mathbf{e}_\alpha$ occurs at a rate of

$$(2.5) \quad \lambda_0 w_\alpha \frac{w_\alpha}{2N} \sum_{\beta(\neq \alpha)} Q_{\alpha\beta} + \lambda_0 \sum_{\beta(\neq \alpha)} w_\beta \frac{w_\alpha}{2N} (1 - Q_{\beta\alpha}),$$

where $Q_{\alpha\beta}$ is a rate at which a haplotype α changes to a haplotype β . For example, $Q_{110,100} = c(1+b)/2 + u(1-u)^2$. We consider the limiting diffusion approximation of this model. We set $\lambda_0 = N$ and assume that $2Nu \rightarrow \theta$ and $2Nc \rightarrow \gamma$ as $N \rightarrow \infty$. $X_i(t)$ denotes the fraction of genes of type A_1 in the i -th locus in the limiting process at time t . We have

$$(2.6) \quad \frac{W_\alpha(t)}{2N} = \prod_{i=1}^n (X_i(t))^{\alpha_i} (1 - X_i(t))^{1-\alpha_i},$$

where α_i is the i -th digit of α . Then, the generator of the diffusion process $\{\mathbf{X}(t); t \geq 0\}$ in an n -dimensional cube $[0, 1]^n$ is exactly identical to that in Eq. 2.3. The limiting diffusion also appears as the limit of the discrete-time Wright-Fisher model.

The probability of fixation of allele A_1 whose initial frequencies \mathbf{p} , $\pi(\mathbf{p})$, satisfies a partial differential equation

$$(2.7) \quad L\pi(\mathbf{p}) = 0$$

with $\theta = 0$ and a boundary conditions $\pi(\mathbf{0}) = 0, \pi(\mathbf{1}) = 1, \pi|_S = \text{finite}$, where $S = \partial[0, 1]^n - \{\mathbf{0}, \mathbf{1}\}$. For details on the boundary conditions, see [16]. We expand the solution as

$$(2.8) \quad \pi(\mathbf{p}) = \pi^{(0)}(\mathbf{p}) + b\pi^{(1)}(\mathbf{p}) + O(b^2).$$

Substituting Eq. 2.8 into Eq. 2.7, we obtain

$$(2.9) \quad \pi^{(0)}(\mathbf{p}) = \bar{p},$$

$$(2.10) \quad \pi^{(1)}(\mathbf{p}) = -\bar{p}\{n - 1 + n\gamma(1 - \bar{p})\} + \frac{2}{n} \sum_{i < j} p_i p_j.$$

It is straightforward to confirm that Eqs. 2.9-2.10 are valid by substituting them into Eq. 2.7 and the boundary conditions; see Appendix for the derivation. We have

$$(2.11) \quad \pi\left(\frac{\mathbf{e}_i}{N}\right) \rightarrow \frac{1}{nN} \{1 - (n - 1)(1 + \gamma')b + O(b^2)\},$$

as $N \rightarrow \infty$, where $\gamma' = n\gamma/(n - 1)$. In terms of the model of gene conversion among members of a multigene family, this expression agrees with that in the weak conversion limit ($\gamma \rightarrow 0$, Eq. 8 in [18]). The effect of bias on the probability of fixation still remains under the weak conversion limit ($\gamma \rightarrow 0$). The weak conversion limit is different from the case that $c = 0$. In the weak conversion limit, all loci are monomorphic except for very short periods of time when a polymorphism is segregating at a single locus. Genetic drift causes the segregating locus to become fixed, either for the introduced allele or for the original allele. After some long length of time, biased conversion creates another polymorphic locus, and the process continues until all loci are fixed for the same allele. Since the locus-by-locus spreading process is biased, the effect of bias should present in the probability of fixation. Note that, the spreading process is impossible for the case that

$c = 0$. Walsh [18] showed that if selection is weak, a slight conversion bias can alter the probability of fixation. When γ is large, the effect could be significant, as shown recently by simulations [16]. Eq. 2.11 shows that under neutrality the deviation of the probability of fixation from that without bias is proportional to one plus the number of converted genes. $\pi(\mathbf{e}_1/N)$ yields the substitution rate. Since time is measured in units of $\lambda_0 = N/2$ birth events, that is

$$(2.12) \quad \frac{nNu}{2} \times \pi(\mathbf{p}_0) = \frac{u}{2} \{1 - (n-1)(1 + \gamma')b + O(b^2)\}.$$

3. A BIASED VOTER MODEL AND THE ANCESTRAL BIAS GRAPH

Krone and Neuhauser showed that the continuous-time Moran model with selection and mutation can be formulated in terms of the biased voter model with mutations on a complete graph [15]. The above introduced continuous-time Moran model with n -demes under biased migration also has an alternative formulation in terms of the biased voter model on a set of complete graphs. Let $\mathbf{I} = (I_i)$, $I_i = \{1, 2, \dots, N\}$, $i = 1, 2, \dots, n$ denote sets of sites, where I_i is the set of sites in the i -th graph. The biased voter model with mutation and biased migration on the set of graphs is a continuous-time Markov process whose state at time t is denoted by $\eta_t : \mathbf{I} \rightarrow \{1, 2\}$. If $x \in I_i$, $\eta_t(x) = 1$ (2), then we say that x is occupied by an individual of type A_1 (A_2) at time t . The process $\{\eta_t; t \geq 0\}$ evolves according to the following rules: (i) For $x = 1, 2, \dots, N$ and $i = 1, 2, \dots, n$, the individual at $x \in I_i$ produces an offspring at rate of λ_0 within I_i ; (ii) The offspring has the same type as the parent with probability $1 - u$ and has the other type with probability u ; (iii) For $x = 1, 2, \dots, N$, $j \neq i$; $i, j = 1, 2, \dots, n$, the individual at $x \in I_i$ produces an offspring in I_j at rates depending on the allelic type. If $\eta_t(x) = 1$ (2), the rate is $\lambda_0\xi_1$ ($\lambda_0\xi_2$); (iv) At the time when the birth event occurs, one of the N sites is chosen at random and the individual at this site is replaced by the offspring. (The offspring is allowed to replace its own parent.). We assume that $\xi_2 - \xi_1 = 2mb/(n-1)$. The birth event in the biased voter model is hierarchical. Two types of birth events, namely, events within a

graph and between graphs, are considered. A bias is present only if a birth event involves different graphs.

The process can be constructed using a percolation diagram [15, 19]. The idea is to construct the process using a collection of independent Poisson processes by drawing arrows on the space-time coordinate system $\mathbf{I} \times [0, \infty)$. These arrows indicate where and when the offspring is produced as well as it is sent. We begin by connecting arrows to each timeline at the times of arrivals in a Poisson process that describes the birth process. For each $(x, y) \in I_i^2$, $i = 1, 2, \dots, n$, let $\{W_{i,s}^{x,y}; s \geq 1\}$ denote the times of arrivals in a Poisson process with rate λ_0/N . For each $(x, y) \in I_i \times I_j$, $i \neq j$, $i, j = 1, 2, \dots, n$, let $\{Z_{i,j,s}^{x,y}; s \geq 1\}$ denote times of arrivals in a Poisson process with rate $\lambda_0\xi_2/N$. Let $\{U_{i,j,s}^{x,y}; s \geq 1\}$ and $\{V_{i,s}^{x,y}; s \geq 1\}$, $i \neq j$, $i, j = 1, 2, \dots, n$ be sequences of independent, uniformly distributed random variables in $(0, 1)$. For times $W_{i,s}^{x,y}$ we draw an arrow from $x \in I_i$ to $y \in I_i$ to indicate the birth of an offspring at x that is sent to y . For times $Z_{i,j,s}^{x,y}$ we draw an arrow from $x \in I_i$ to $y \in I_j$ to indicate the birth of an offspring at x which is then sent to y . If $U_{i,j,s}^{x,y} < \xi_1/\xi_2$, we place a “ δ ” at the tip of the arrow; otherwise, we label the arrow with a “2”. In other words, we have δ -arrows and 2-arrows entering a site y at rates $\lambda_0\xi_1$ and $\lambda_0(\xi_2 - \xi_1)$, respectively. Then, the following rule will apply: 2’s can give birth through both types of arrows, but 1’s can only give birth through δ -arrows. The process $\{V_{i,s}^{x,y}; s \geq 1\}$ is used as the mutation process; if $V_{i,s}^{x,y} < u$, a mutation occurs. We represent a mutation event by solid dots on the arrows. A realization of the percolation diagram in the case $n = 2$ and $N = 4$ is shown in Fig. 1. I_1 and I_2 are the left and the right graph, respectively. If the set of 1’s initially is $\{1\} \in I_2$, then, at time t , the set of 1’s is $\{2\} \in I_1$ and $\{3\} \in I_2$. The paths of the 1’s are indicated by thick lines. By reversing time, we can follow the ancestral history of individuals at sites in a finite set and thus determine their types. The resulting process is called the dual or ancestral process [10, 15, 20]. A realization of the dual process, which was obtained from Fig. 1 by simply reversing time and the direction of arrows, is shown in Fig. 2. Here, the ancestral history

of a sample consists of individuals at sites $\{1, 2\} \in I_1$, and $\{1\} \in I_2$ at dual time 0 is indicated by thick lines.

To obtain the analogue of the coalescent, we rescale time and the parameters as in the previous section: $\lambda_0 = N/2, \xi_2 = m(1+b)/(n-1), \xi_1 = m(1-b)/(n-1)$ with $Nu \rightarrow \theta$ and $Nm \rightarrow \gamma$ as $N \rightarrow \infty$. The dynamics of the dual process follow readily from the percolation diagram. We can ignore the event in which a particle in the dual process crosses an unmarked arrow and lands either on a site not present in the dual process or on its own site. To describe the other possible events, assume that there are $\mathbf{k} = (k_i), i = 1, 2, \dots, n$ particles in the dual process, where k_i is the number of particles contained in I_i . We say that a coalescing event has occurred when a particle crosses an unmarked arrow and lands on the site of different particle contained in the dual process. This occurs at rates

$$(3.1) \quad \lambda_0 k_i \frac{k_i - 1}{N} = \frac{k_i(k_i - 1)}{2}, \quad i = 1, 2, \dots, n.$$

We say that a migration event has occurred when a particle crosses a δ -arrow. This occurs at rates

$$(3.2) \quad \lambda_0 \xi_1 k_i (N - k_j) \rightarrow \frac{(1-b)\gamma}{2(n-1)} k_i, \quad N \rightarrow \infty,$$

for $j \neq i; i, j = 1, 2, \dots, n$. We say that branching event has occurred when a particle crosses a 2-arrow. This occurs at a rate of

$$(3.3) \quad \lambda_0 (\xi_2 - \xi_1) k_i (N - k_j) \rightarrow \frac{b\gamma}{n-1} k_i, \quad N \rightarrow \infty,$$

for $i \neq j; i, j = 1, 2, \dots, n$. The original particle continues along the old path (continuing branch) and the new particle that arose from the branching follows the 2-arrow (incoming branch). If the new particle lands on a site that is already contained in the dual process, we say a collision has occurred, but collisions can be ignored in the diffusive limit $N \rightarrow \infty$. We call the random graph generated by the dual process in the diffusive limit as the ancestral bias graph.

Let $\mathbf{A}_n(t) = (A_{n,i}(t))$, $i = 1, 2, \dots, n$ denote the number of particles that are present in the dual time at $t \geq 0$, where $A_{n,i}(t)$ is the number of particles contained in I_i . We set $\mathbf{A}_n(0) = \mathbf{n}$. The size process $\mathbf{A}_n(t)$ is an n -dimensional birth and death process. If $\mathbf{A}_n(t) = \mathbf{k}$, the transition to $\mathbf{k} + \mathbf{e}_i$ is at a rate $b\gamma(n\bar{k} - k_i)/(n - 1)$ and the transition to $\mathbf{k} - \mathbf{e}_i$ is at a rate at $k_i(k_i - 1)/2$. Due to the moment duality between the birth and death process and the Wright-Fisher diffusion governed by the generator Eq. 2.3, the stationary measure of the birth and death process can be obtained by the probability of fixation in the Wright-Fisher diffusion Eq. 2.8, as was shown in the ancestral selection graph [21]. In particular

$$(3.4) \quad \phi(\mathbf{e}_i) = \frac{1}{n} - b \left(\frac{n-1}{n} - \gamma \right) + O(b^2),$$

$$(3.5) \quad \phi(2\mathbf{e}_i) = b\frac{\gamma}{n} + O(b^2),$$

$$(3.6) \quad \phi(\mathbf{e}_i + \mathbf{e}_j) = b\frac{2}{n}(1 + \gamma) + O(b^2),$$

for $i \neq j$; $i, j = 1, 2, \dots, n$. Other configurations are $O(b^2)$. In the weak migration limit ($\gamma \rightarrow 0$), the configurations that have non-zero probabilities up to $O(b^2)$ are \mathbf{e}_i and $\mathbf{e}_i + \mathbf{e}_j$. In the weak migration limit, compared with waiting times for migration and branching events, a waiting time for coalescing event is negligible. Thus, the state should be whether an ancestor is in a deme or two ancestors are in different demes, where one of which was produced by a recent branching event, and they are waiting for being in the same deme and coalescing. To simulate a sample, we may stop the process at the time at which the total size of the dual process reaches 1, since the types of the particles present at this time determines the types in the sample. We call the particle at this point in time the ultimate ancestor. As for the ancestral selection graph [15], branches in the ancestral bias graph do not necessary represent the true genealogy. Depending on the type of the ultimate ancestor and the mutation events along the branches, certain parts of the ancestral graph may not be accessible to individuals since only individuals of type A_2 may cross 2-arrows. By following the path in the backward direction to the ultimate ancestor, we obtain the

ancestral paths of each individual and hence the true genealogy of the sample. The true genealogy depends on the type of the ultimate ancestor. In Fig. 2, if the type of the ultimate ancestor is A_1 , the true genealogy contains the dotted line and does not contain the dashed line, and vice versa.

Griffiths and Tavaré [22] introduced an importance sampling algorithm for computing the probability distribution for samples taken from a population that evolves according to certain neutral models. The algorithm is based on a recursion satisfied by the sampling distribution. A scheme using Markov chain Monte Carlo to simulate backwards along the sample path of the ancestral bias graph can approximate the sampling distribution of our model by use of similar recursions for the neutral cases. If $\mathbf{f} = (f_i)$, $i = 1, 2, \dots, n$ genes are taken from the i -th deme, where a_i and d_i genes are of type A_1 and of type A_2 , respectively, we say that the sample is of type configuration (\mathbf{a}, \mathbf{d}) . Let \mathbf{X} be distributed according to the stationary distribution. We denote by $q(\mathbf{a}, \mathbf{d})$ the probability that a sample of \mathbf{f} genes taken from a population in equilibrium is of type configuration (\mathbf{a}, \mathbf{d}) . Then, it follows that

$$(3.7) \quad q(\mathbf{a}, \mathbf{d}) = \prod_{i=1}^n \frac{(f_i)!}{a_i! d_i!} \mathbb{E}_{\mathbf{f}} [X_i^{a_i} (1 - X_i)^{d_i}].$$

The probability $q(\mathbf{a}, \mathbf{d})$ satisfies a recursion.

Theorem 3.1. *The probability $q(\mathbf{a}, \mathbf{d})$ satisfies*

$$(3.8) \quad \begin{aligned} r(\mathbf{a}, \mathbf{d}) q(\mathbf{a}, \mathbf{d}) &= \sum_{i=1}^n \{ (a_i - 1) f_i q(\mathbf{a} - \mathbf{e}_i, \mathbf{d}) + (d_i - 1) f_i q(\mathbf{a}, \mathbf{d} - \mathbf{e}_i) \} \\ &+ \theta \sum_{i=1}^n \{ (d_i + 1) q(\mathbf{a} - \mathbf{e}_i, \mathbf{d} + \mathbf{e}_i) + (a_i + 1) q(\mathbf{a} + \mathbf{e}_i, \mathbf{d} - \mathbf{e}_i) \} \\ &+ \frac{(1-b)\gamma}{n-1} \sum_{j \neq i} f_i \left\{ \frac{a_j + 1}{f_j + 1} q(\mathbf{a} - \mathbf{e}_i + \mathbf{e}_j, \mathbf{d}) + \frac{d_j + 1}{f_j + 1} q(\mathbf{a}, \mathbf{d} - \mathbf{e}_i + \mathbf{e}_j) \right\} \\ &+ \frac{2b\gamma}{n-1} \sum_{i \neq j} \left\{ \frac{d_i(d_j + 1)}{f_j + 1} q(\mathbf{a}, \mathbf{d} + \mathbf{e}_j) + \frac{f_i(a_j + 1)}{f_j + 1} q(\mathbf{a} + \mathbf{e}_j, \mathbf{d}) \right. \\ &\left. + \frac{(a_i + 1)(d_j + 1)}{f_j + 1} q(\mathbf{a} + \mathbf{e}_i, \mathbf{d} - \mathbf{e}_i + \mathbf{e}_j) \right\}, \end{aligned}$$

where

$$(3.9) \quad r(\mathbf{a}, \mathbf{d}) = \sum_{i=1}^n f_i \{(f_i - 1) + \theta + \gamma(1 + b)\}.$$

The probabilities with negative arguments are zero. The boundary conditions are

$$(3.10) \quad q(\mathbf{e}_i, \mathbf{0}) = \rho, \quad q(\mathbf{0}, \mathbf{e}_i) = 1 - \rho, \quad i = 1, 2, \dots, n$$

where ρ is the probability that the ultimate ancestor is of type A_1 .

As for Theorem 5.2 in [15] for the ancestral selection graph, since it is straightforward to prove Theorem 3.5, we do not present the proof. Theorem 3.5 can be proved either by computing the moments or using the structure of the ancestral bias graph. ρ is the expected value of relative frequency of A_1 in the diffusive limit under stationarity. Although an analytic expression for ρ is not available, it is possible to simulate it by using coupling from the past [24, 25].

In principle, we can obtain probability of any sample by using the importance sampling algorithm [22], but we focus on analytic expression for a sample of size two ($|\mathbf{f}| = 2$). By solving Eq. 3.8, we have

$$(3.11) \quad q(2\mathbf{e}_i, \mathbf{0}) = \frac{\gamma'/n + \theta + (1 + 2\theta + \gamma')(\theta - b\gamma)}{2\{\gamma'/n + 2\theta(1 + 2\theta + \gamma')\}} + O(b^2),$$

$$(3.12) \quad \begin{aligned} q(\mathbf{e}_i + \mathbf{e}_j, \mathbf{0}) &= \frac{\gamma'/n + (1 + 2\theta + \gamma')(\theta - b\gamma)}{2\{\gamma'/n + 2\theta(1 + 2\theta + \gamma')\}} + O(b^2), \\ q(\mathbf{e}_i, \mathbf{e}_i) &= \frac{\theta(2\theta + \gamma')}{\gamma'/n + 2\theta(1 + 2\theta + \gamma')} + O(b^2). \end{aligned}$$

$$(3.13) \quad q(\mathbf{e}_i, \mathbf{e}_j) = \frac{\theta(1 + 2\theta + n\gamma')}{2\{\gamma'/n + 2\theta(1 + 2\theta + \gamma')\}} + O(b^2),$$

for $i \neq j$; $i, j = 1, 2, \dots, n$. $q(\mathbf{0}, 2\mathbf{e}_i)$, $q(\mathbf{0}, \mathbf{e}_i + \mathbf{e}_j)$ are given by Eqs. 3.11-3.12, respectively, by replacing b by $(-b)$. It can be seen that the effect of conversion bias on the sampling distribution is not as large as the effect of bias on the probability of fixation Eq. 2.11; in the formula for the probability of fixation, the ratio of the correction term in $O(b)$ to

the term of $O(1)$ is proportional to the number of demes/loci. The sampling distribution reduces to that in the model without bias when mutation events dominate biased migration/conversion events ($bc \ll u$) as well as in including the weak migration/conversion limit. Ohta [23] defined the identity coefficients between members of multigene family. f is the average probability of allelic identity, c_1 is the average probability of identity at different loci on one chromosome, and c_2 is that of two genes taken from different loci of two homologous chromosomes. For unlinked loci, $c_1 = c_2$. In terms of the island model, f is the average probability of gene identity for genes sampled from the same deme, while $c_1 = c_2$ is that of two genes taken from different demes. We have

$$(3.14) \quad f = q(2\mathbf{e}_i, \mathbf{0}) + q(\mathbf{0}, 2\mathbf{e}_i),$$

$$(3.15) \quad c_1 = c_2 = q(\mathbf{e}_i + \mathbf{e}_j, \mathbf{0}) + q(\mathbf{0}, \mathbf{e}_i + \mathbf{e}_j).$$

When $b = 0$, Eqs. 3.14-3.15 reduce to Eqs. 12 in [23].

4. THE STRONG MIGRATION/CONVERSION LIMIT

Nagylaki [14] established the strong-migration limit for a geographically structured population. Let $X(t) = \sum_{i=1}^n Z_i(t)/(nN)$, which is the frequency of A_1 in the entire population. Since the deme size is not altered in our n -island model, the effective size is the total size, and all effects of population subdivision disappear [14]. It is straightforward to verify the conditions (Eqs. 22) in [14] are satisfied, and we obtain the limiting diffusion of the continuous-time Markov chain $\mathbf{Z}(t)$ with m and n fixed, $Nb \rightarrow \beta$ as $N \rightarrow \infty$ (strong migration limit) in the diffusion time units of n . The generator is

$$(4.1) \quad L = \frac{x(1-x)}{2} \frac{\partial^2}{\partial x^2} - \left\{ nm\beta x(1-x) - \frac{n\theta}{2}(1-2x) \right\} \frac{\partial}{\partial x}.$$

Eq. 4.1 is identical to that of the diffusion for mutation and selection, which has been extensively investigated; here, the mutation rate is nu and the selection intensity is $2nmb$.

The probability of fixation of allele A_1 , whose initial frequencies are \mathbf{p} , is given by

$$(4.2) \quad \pi(\mathbf{p}) = \frac{e^{2nm\beta\bar{p}} - 1}{e^{2nm\beta} - 1},$$

which agrees with Eq. 2.8. In the strong migration limit, the ancestral bias graph reduces to the ancestral selection graph, which has been extensively investigated [15, 21], with the scaled selection intensity (σ in [15]) is $2nm\beta$. In the strong migration limit, the population becomes panmictic and branching events are dominate. The branching rate for each individual in the ancestral bias graph is $nm\beta$ (note that time is measured in units of n), which is equal to the rate in the ancestral selection graph whose selection intensity is $2nm\beta$. By using the density of the unique stationary distribution (Wright's formula [26]), it is possible to obtain an analytic expression of $q(\mathbf{a}, \mathbf{d})$.

5. BIAS IN ECTOPIC GENE CONVERSION AMONG MULTIGENE FAMILY

An exon sequence (393 base pairs) of mouse histone H2A gene (single exon gene) is retrieved from Ensembl release 53 (<http://www.ensembl.org/>) and hypothetical family member genes are searched from the complete mouse genome build 37.1 with using BLASTN [27]. A minimum of 90% similarity to the reference sequence and 90% coverage of the family member genes were required. We obtain 36 sequences, whose GC content at the third codon position was 88.0%, which is significantly higher than average GC content in the mouse genome (42%) [28]. This high GC content is probably due to biased gene conversion among the histone gene family [6]. As long as the bias b is small, the moment estimates of the parameters by using the formula for the identity coefficients Eq. 3.15 without bias and that for the substitution rate Eq. 2.12 have an accuracy up to $O(b^2)$. The actual process of conversion is likely to involve a piece of a gene. In the analysis below, a nucleotide site is considered. Here, c is the average rate at which the nucleotide converted by the homologous nucleotide of another locus belonging to the multigene family. We assume $n = 36$ and the substitution rate is 1.22×10^{-9} per site per generation. The substitution

rate was estimated by noting the sequence divergence between the mouse sequence and the homologous rat sequence at the third codon position (14.5%), that estimated mouse-rat divergence time (33million years [29]), and the average generation time of the rat (0.5years). We dichotomize nucleotide to AT/GC and set AT and GC nucleotides as allele A_1 and A_2 , respectively. Under the assumption that the substitutions occurs symmetrically, 2/3 of the substitution occur between AT and GC and we have $u = 0.813 \times 10^{-9}$. Assuming rat effective size as $N = 1.61 \times 10^5$ [30], we have $\theta = 2.62 \times 10^{-4}$. Average identity coefficient at the third codon position in comparison between two sequences was $c_2 = 0.840$. According to Eq. 3.15, we estimate $c = 1.26 \times 10^{-7}$. (The estimate is robust to the dichotomy of nucleotides and the assumption of free recombination. Even if we use a formula for complete linkage with four allele model [23], we have $c = 1.00 \times 10^{-7}$). From the substitution rate Eq. 2.12, the expected fraction of genes of type A_1 at a site is

$$(5.1) \quad \eta(t) = \tilde{\eta} + (\eta(0) - \tilde{\eta})e^{-ut}, \quad \tilde{\eta} = \frac{1 - (n-1)(1 + \gamma')b}{2}.$$

If the GC content reaches its equilibrium, $\tilde{\eta} = 0.880$ and we have $b = 0.0209$. Backstrom et al. [31] showed that chicken HINTW gene family linked to the W chromosome seems to undergo gene conversion at a rate of $c = (3-4) \times 10^{-6}$. They reported that the GC content of the intron (55.3%) is significantly higher than that of introns of other W-linked genes (40.0%). If the GC content reaches the equilibrium, we have $b = (6-8) \times 10^{-4}$. If the GC content does not reach the equilibrium, the actual bias could be larger. Recent estimates of scaled allelic conversion bias ($4Nb'$, where the GC gametes produced by a heterozygote individual being given by $(1 + b')/2$ [3]) in high-recombination region are 1.7 or 0.60 in *Drosophila* [32, 33], and 12.7 in humans [34]. It is suggested that ectopic gene conversion is also biased, whose magnitude could be orders of magnitude larger than that in allelic gene conversion. The relatively strong bias in ectopic gene conversion is plausible; bias in allelic gene conversion is equivalent to genic selection whose intensity is b , while bias in ectopic gene conversion is equivalent to genic selection whose intensity is $2ncb$, at least

in the strong migration limit Eq.4.1. For the mouse histone H2A gene family, the scaled conversion bias is $2nc\beta = 0.122$ and the effect of the bias is almost neutral. Since nc is sufficiently smaller than unity except for very large multigene family, relatively large conversion bias can be maintained in a population against purifying selection.

6. DISCUSSION

It was recently found that the diffusion models of genic selection in an n -island model and that of gene conversion among members of a multigene family in a panmictic population are identical [16]. In this study, we have developed two continuous-time Moran models of bias in evolutionary mechanisms: an n -island model and a model of ectopic gene conversion among members of a multigene family in a panmictic population. The models are very different but have the same diffusive limit, where allele frequencies in each subpopulation of the n -island model are identical to allele frequencies in each locus of n -gene family. Nagylaki [3] showed that bias in allelic gene conversion is equivalent to genic selection. In contrast, it was shown that bias in ectopic gene conversion is equivalent to genic selection only in the strong conversion limit. The n -island model was formulated in terms of a hierarchical biased voter model, and the ancestral bias graph was introduced as an object generated by the dual process. As for the ancestral selection graph [15], it is possible to compute the quantities of interest, such as the probability of genes being identical by descent and the time to the most recent common ancestor, by expanding the ancestral bias graph by the population-scaled migration/conversion rate (γ). Due to the moment duality between the diffusion process and the biased voter model in the diffusive limit, we can study some properties of the ancestral bias graph from the probability of fixation in the diffusion process (See [21]), and the reduction of the ancestral bias graph to the ancestral selection graph was shown via the strong migration/conversion limit of the diffusion process. The model presented here is minimal, but the migration and conversion schemes can be extended in various ways.

In this study, we have investigated concerted evolution by biased gene conversion among members of a multigene family. Unequal crossing-over is the other possible mechanism of the converted evolution. Unequal crossing-over occurs fairly frequently between nonallelic homologous genes in a large multigene family, where large number (the order of 100 or even 1,000) of homologous genes are tandemly arranged on the chromosome [35, 36]. When average shift of gene units is large, unequal crossing-over can be dominate [35]. Bias in gene conversion is considered to be caused by bias in repair of doublestrand breaks by recombinations. Since concerted evolution by unequal crossing-over does not involve the bias, unequal crossing-over is probably less biased. In addition, bias in concerted evolution of large multigene family will be difficult to evolve in a population, even if biased gene conversion operates. Since the scaled conversion bias is proportional to the number of loci, bias in large gene family will be deleterious and pushed out from a population by purifying selection.

Remarkably, bias in gene conversion and that in migration are mathematically equivalent phenomena in the diffusive limit. Recently, evidence of the large impact of biased gene conversion on gene substitution is accumulating. Although the data that demonstrate migration bias are still limited, we can speculate that there might be a slight bias in the migration rate associates with alleles. If the number of migrant is large, effect of migration bias is equivalent to genic selection. Since a gene conversion bias of a few percent could cause a substantial increase in the GC content, it seems likely that slight bias in migration has large impact on population differentiation and speciation in natural populations. If population dispersal involved gradual population subdivision, the effective size could have been reduced without a reduction in the census population size; the effective size could be reduced substantially by rapid fixation of alleles associate to people who migrate quickly.

Acknowledgments

I thank B. Charlesworth (University of Edinburgh, U.K.), who suggest the subject of this paper to me. I thank B. Charlesworth, T. Ohta (National Institute of Genetics, Japan), K. Zeng (University of Edinburgh, U.K.) for comments on the manuscript.

7. APPENDIX

The diffusion process $\{\mathbf{X}(t); t \geq 0\}$ in an n -dimensional cube $[0, 1]^n$ has exit boundaries at $\mathbf{0}$ and $\mathbf{1}$. As $t \rightarrow \infty$ finite probability mass remains only at these exit boundaries, and we have

$$(7.1) \quad \lim_{t \rightarrow \infty} \mathbb{E}[X_i(t)] = \pi(\mathbf{p}), \quad i = 1, 2, \dots, n.$$

Let $\mu_{\mathbf{a}}(t) = \mathbb{E}[\prod_{i=1}^n X_i^{a_i}(t)]$ and expand the Laplace transform by a power series in b

$$(7.2) \quad \nu_{\mathbf{a}}(s) = \nu_{\mathbf{a}}^{(0)}(s) + b\nu_{\mathbf{a}}^{(1)}(s) + O(b^2).$$

At the 0-th order in b we have a system of equations

$$(7.3) \quad (s + \gamma')\nu_{\mathbf{e}_i}^{(0)} - \frac{\gamma'}{n} \sum_{j=1}^n \nu_{\mathbf{e}_j}^{(0)} = p_i, \quad i = 1, 2, \dots, n.$$

The solution is

$$(7.4) \quad \nu_{\mathbf{e}_i}^{(0)}(s) = \frac{\bar{p}}{s} + \frac{p_i - \bar{p}}{s + \gamma'}, \quad i = 1, 2, \dots, n.$$

Consult [16] for the derivation. By applying the inverse Laplace transform, we have $\mu_{\mathbf{e}_i}^{(0)} = \bar{p} + (p_i - \bar{p})e^{-\gamma' t}$. Thus, $\pi^{(0)}(\mathbf{p}) = \bar{p}$. In the same manner, for $i = 1, 2, \dots, n$,

$$(7.5) \quad \left(\frac{s}{2} + 1 + \gamma'\right) \nu_{2\mathbf{e}_i}^{(0)} - \nu_{\mathbf{e}_i}^{(0)} - \frac{\gamma'}{n} \sum_{j=1}^n \nu_{\mathbf{e}_i + \mathbf{e}_j}^{(0)} = \frac{p_i^2}{2}$$

and for $i \neq j$; $i, j = 1, 2, \dots, n$,

$$(7.6) \quad (s + 2\gamma')\nu_{\mathbf{e}_i + \mathbf{e}_j}^{(0)} - \frac{\gamma'}{n} \sum_{k=1}^n (\nu_{\mathbf{e}_j + \mathbf{e}_k}^{(0)} + \nu_{\mathbf{e}_i + \mathbf{e}_k}^{(0)}) = p_i p_j.$$

They can be solved for $\nu_{2\mathbf{e}_i}^{(0)}$ and $\nu_{\mathbf{e}_i+\mathbf{e}_j}^{(0)}$. At the 1-st order in b , we have a system of equations for $i = 1, 2, \dots, n$,

$$(7.7) \quad \begin{aligned} & (s + \gamma')\nu_{\mathbf{e}_i}^{(1)} - \frac{\gamma'}{n} \sum_{j=1}^n \nu_{\mathbf{e}_j}^{(1)} \\ &= \frac{\gamma'}{n} \left\{ (n-2)\nu_{\mathbf{e}_i}^{(0)} + \sum_{j=1}^n \nu_{\mathbf{e}_j}^{(0)} - 2 \sum_{j(\neq i)} \nu_{\mathbf{e}_i+\mathbf{e}_j}^{(0)} \right\}. \end{aligned}$$

Substituting Eq. 7.4 and solutions for Eqs. 7.5-7.6 into Eq. 7.7, we have

$$(7.8) \quad \nu_{\mathbf{e}_i}^{(1)}(s) = \frac{a_0}{s} + \sum_{j=1}^n \frac{a_j}{s - s_j}, \quad i = 1, 2, \dots, n,$$

where

$$(7.9) \quad a_0 = (n-1)\bar{p}\{1 + \gamma'(1 - \bar{p})\} - \frac{2}{n} \sum_{i < j} p_i p_j.$$

s_j are eigenvalues of the generator Eq. 2.3 and a_j , $j \neq 0$ are constants which do not depend on s . Then, by applying the inverse Laplace transform, Eq. 2.10 follows.

REFERENCES

- [1] Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159** 907–911.
- [2] Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genet* **23** 273–277.
- [3] Nagylaki T (1983) Evolution of a finite population under gene conversion. *Proc Natl Acad Sci USA* **80**: 6278–6281.
- [4] Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol* **7** e1000026.
- [5] Ohta T (1980) Evolution and variation of multigene families. Springer-Verlag, Berlin/New York.
- [6] Galtier N (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends Genet* **19** 65–68.
- [7] Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat Genet* **29** 20–21.

- [8] Ségurel L et al. (2008) Sex-specific genetic structure and social organization in central asia: insights from a multi-locus study. *PLoS Genet* **4** e1000200.
- [9] Wright S (1951) The genetical structure of populations. *Ann Eugen* **15** 323–354.
- [10] Harris TE (1976) On a class of set-values Markov process. *Ann Probab* **4** 175–194.
- [11] Kingman JFC (1982) The coalescent. *Stoch Proc Appl* **13** 235–248.
- [12] Tajima F (1983) Evolutionary relationship of DNA sequences in finite population. *Genetics* **105** 437–460.
- [13] Hudson RR (1983) Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37** 203–217.
- [14] Nagylaki T (1980) The strong migration limit in geographically structured populations. *J Math Biol* **9** 101–114.
- [15] Krone SM, Neuhauser C (1997) Ancestral process with selection. *Theor Popul Biol* **51** 210–237.
- [16] Mano S, Innan H (2008) The evolutionary rate of duplicated genes under concerted evolution. *Genetics* **181** 493–505.
- [17] Nagylaki T, Petes T (1982) Interchromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics* **100** 315–337.
- [18] Walsh JB (1985) Interaction of selection and biased gene conversion in a multigene family. *Proc Natl Acad Sci USA* **82** 153–157.
- [19] Harris TE (1972) Nearest neighbor Markov interaction process on multidimensional lattices. *Adv Math* **9** 66–89.
- [20] Donnelly P (1984) The transient behavior of the Moran model in population genetics. *Proc Camb Phil Soc* **95**: 349–358.
- [21] Mano S (2009) Duality, ancestral and diffusion processes in models with selection. *Theor Popul Biol* **75** 164–175.
- [22] Griffiths RC, Tavaré S (1994) Simulating probability distribution in the coalescent. *Theor Popul Biol* **46** 131–159.
- [23] Ohta T (1982) Allelic and nonallelic homology of a supergene family. *Proc Natl Acad Sci USA* **79** 3251–3254.
- [24] Propp JG, Wilson DB (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct Algorithms* **9** 223–252.
- [25] Fearnhead P (2001) Perfect simulation from population genetic models with selection. *Theor Popul Biol* **59** 263–279.

- [26] Wright S (1938) The distribution of gene frequencies in population of polyploids. *Proc Natl Acad Sci* **24** 372–377.
- [27] Altschul SF et al. (1990) Basic local alignment search tool. *J Mol Biol* **215** 403–410.
- [28] Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420** 520–562.
- [29] Nei M, Xu P, Glazko G (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc Natl Acad Sci USA* **98** 2497–2502.
- [30] Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D (2002) Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol* **19** 2142–2149.
- [31] Backström N, Cepitis H, Berlin S, Ellegren H (2005) Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. *Mol Biol Evol* **22** 1992–1999.
- [32] Galtier N, Bazin E, Bierne N (2006) GC-biased segregation of noncoding polymorphism in *Drosophila*. *Genetics* **172** 221–228.
- [33] Haddrill P, Charlesworth B (2008) Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biol Lett* **4** 438–441.
- [34] Lercher MJ, Smith NGC, Eyre-Walker A, Hurst LD (2002) The evolution of isocores: evidence from SNP frequency distributions. *Genetics* **162** 1805–1810.
- [35] Ohta T (1983) On the evolution of multigene families. *Theor Popul Biol* **23** 211–235.
- [36] Eickbush TH, Eickbush DG (2007) Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* **175** 477–485.

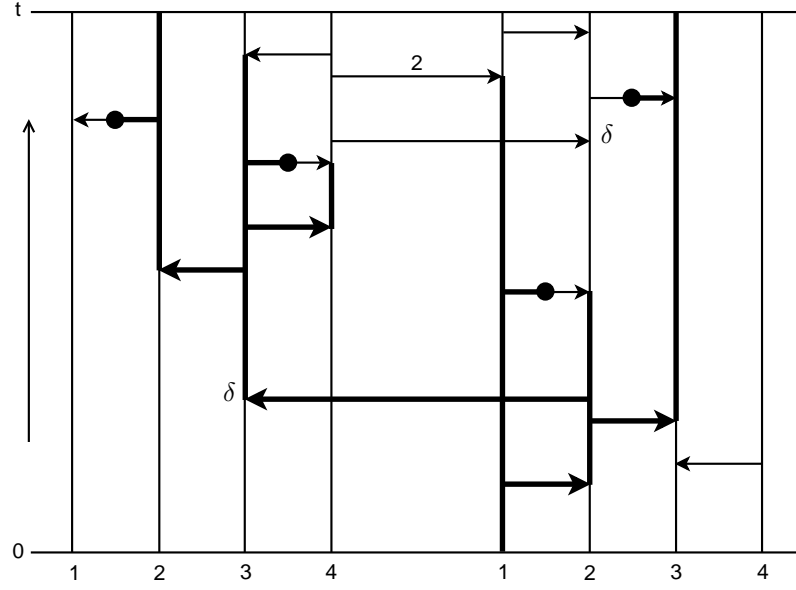


FIGURE 1. The graphical representation for the biased voter model for the case $n = 2$ and $N = 4$. If initially the set of A_1 's is $\{1\} \in I_2$, then at time t , the set of A_1 's is $\{2\} \in I_1$ and $\{3\} \in I_2$. I_1 and I_2 are the left and the right graph, respectively. The paths of A_1 's are indicated by thick lines.

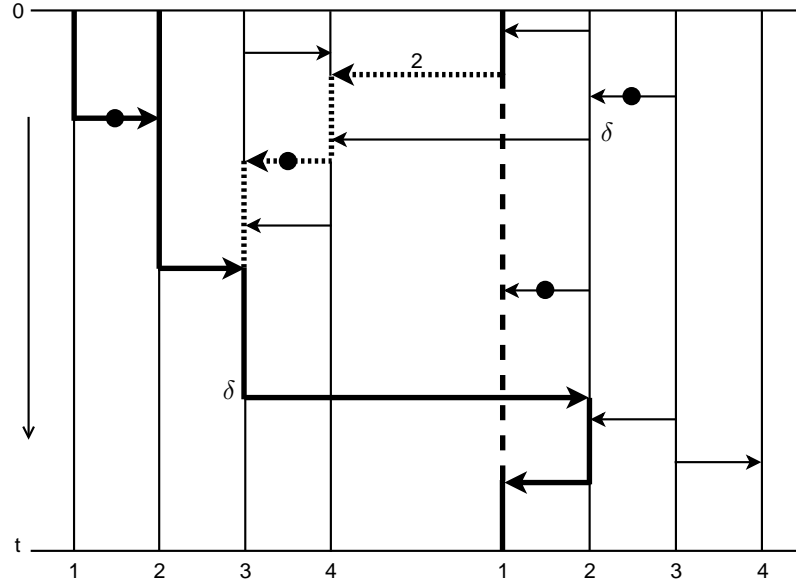


FIGURE 2. The graphical representation for the dual process of the biased voter model. Here, the ancestral history of a sample, consists of individuals at sites $\{1, 2\} \in I_1$ and $\{1\} \in I_2$ at dual time 0 is indicated by thick lines. The ultimate ancestor is in $\{1\} \in I_2$, at dual time t . If the ultimate ancestor is A_1 , then the true genealogy contains the dotted line and does not contain the dashed line, and vice versa.